DOCUMENT RESUME

ED 213 736 TM 820 124

AUTHOR Wise, Lauress L.; McLaughlin, Donald H.
TITLE Survey Data Enhancement.
PUB DATE Apr 81
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Analysis; *Editing; *Research Methodology; *Surveys
IDENTIFIERS Data Editing; Longitudinal Merges; *Missing Data

ABSTRACT

Work performed during the 1978-1980 SAGE contract to develop improved national estimates from survey data is reported. Three areas of effort are covered in this paper: (1) the use of longitudinal merges combined with relational edits to detect reporting or encoding errors; (2) the use of longitudinal merges together with special follow-up surveys to improve the universe coverage; and (3) the use of missing data imputation techniques to develop national estimates when key data elements are missing due to nonresponse or omissions. (Author/GK)

SURVEY DATA ENHANCEMENT*

Lauress L. Wise
Donald H. McLaughlin

American Institutes for Research

This paper reports on work performed during the 1978-1980 SAGE contract to develop improved national estimates from survey data. Three areas of effort are covered in this paper. The first is the use of longitudinal merges combined with relational edits to detect reporting or encoding errors. The second is the use of longitudinal merges together with special follow-up surveys to improve the universe coverage; and the third is the use of missing data imputation techniques to develop national estimates when key data elements are missing due to nonresponse or omissions.

RELATIONAL EDITS

The first area of SAGE work to be discussed here was the development of edit specifications for data from the Common Core of Data (CCD). In particular, parts VI and VIa of this data base include data on each of the nation's public school districts (LEAs) and on each public school. While the number of data elements for each LEA or school is small, the very large number of units in each file makes it a virtual certainty that data reporting and/or data entry errors will creep into the file. An important way in which survey data such as these can be enhanced is to find and correct such erroneous values.

An efficient edit procedure must identify a high proportion of the invalid responses while not also flagging so many valid responses as to make checking each identified case infeasible. In the absence of any other information, the traditional procedure is to examine the most extreme values, both because these are least likely to be valid and

---

* Paper presented at the 1981 Annual Meeting of the American Educational Research Association, April, 1981

because they have the greatest impact on summary statistics. Unfortunately, the range of valid values for these CCD files is so great that such an edit would be meaningless. If a district served 800 pupils, but 8,000 was erroneously entered, for example, there would be little chance of catching this error with a simple range check. The value of 8,000 is perfectly valid for many districts.

The relational editing strategy proposed by SAGE uses values that are closely correlated with each field being edited to "predict" the value in question and then compares the actual values with these predictions. The greatest discrepancies are flagged for further checking. In the example cited, the error in the number of students might have been caught because it led to to an unreasonable ratio of students to teachers, or of students to schools, in the district.

By far the best predictor of any of the values in the LEA and Public School surveys is the corresponding value from the prior year's survey. Therefore, longitudinal merges were proposed to allow the comparison of values between successive years. To illustrate the effectiveness of this approach, some data were taken from NCES's Nonpublic School Surveys. Figure 1 shows the distribution of the number of pupils served by each nonpublic school. This distribution is very broad. If we wanted to examine only schools with the most extreme values, say the upper and lower 1%, we would have to accept all values between about 5 and 1100. Figure 2 shows the distribution of the differences between the 1977-78 values and the corresponding values from the 1976-77 survey. In this case the range of values accepted without question would be only about 200 (-100 to 100) rather than 1100. Most kinds of recording or data entry errors are relatively infrequent and random so that the probability that both the new and the prior values contain compensating errors is negible. In this case, virtually all errors of any significant magnitude would be flagged while few valid responses would be flagged.

Figure 2 also shows that the difference values have a nearly normal distribution, particularly in comparison to the highly skewed distribution in Figure 1. To the extent that the true values do follow a normal distribution, we have some basis for estimating the proportion of "error" values above or below any given cutoff by comparing the actual distribution with the predicted distribution. Figure 2 shows a normal distribution superimposed over the actual difference distribution. The relatively

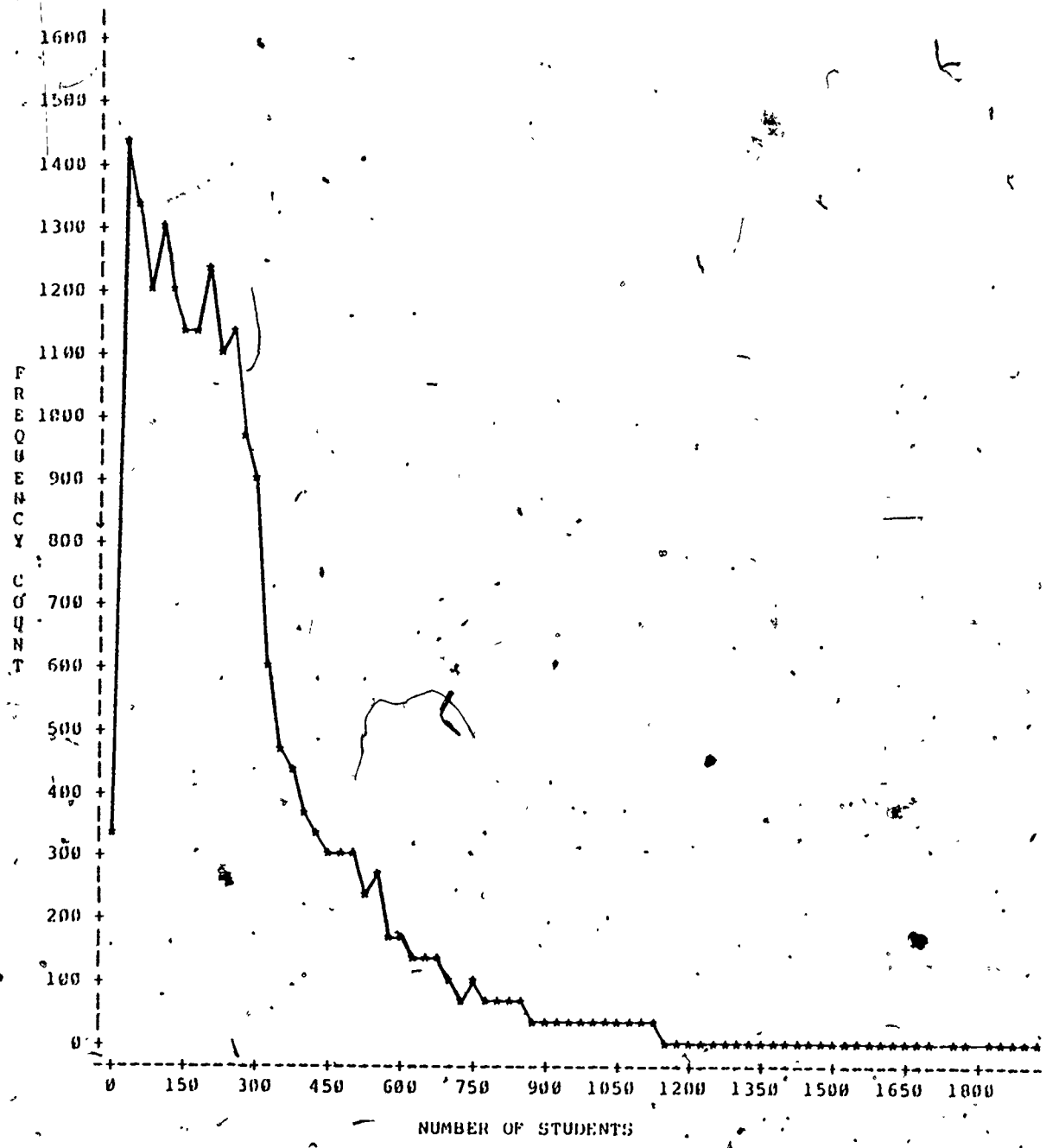

FIGURE 1. DISTRIBUTION OF 1977 ENROLLMENT OF NONPUBLIC SCHOOLS

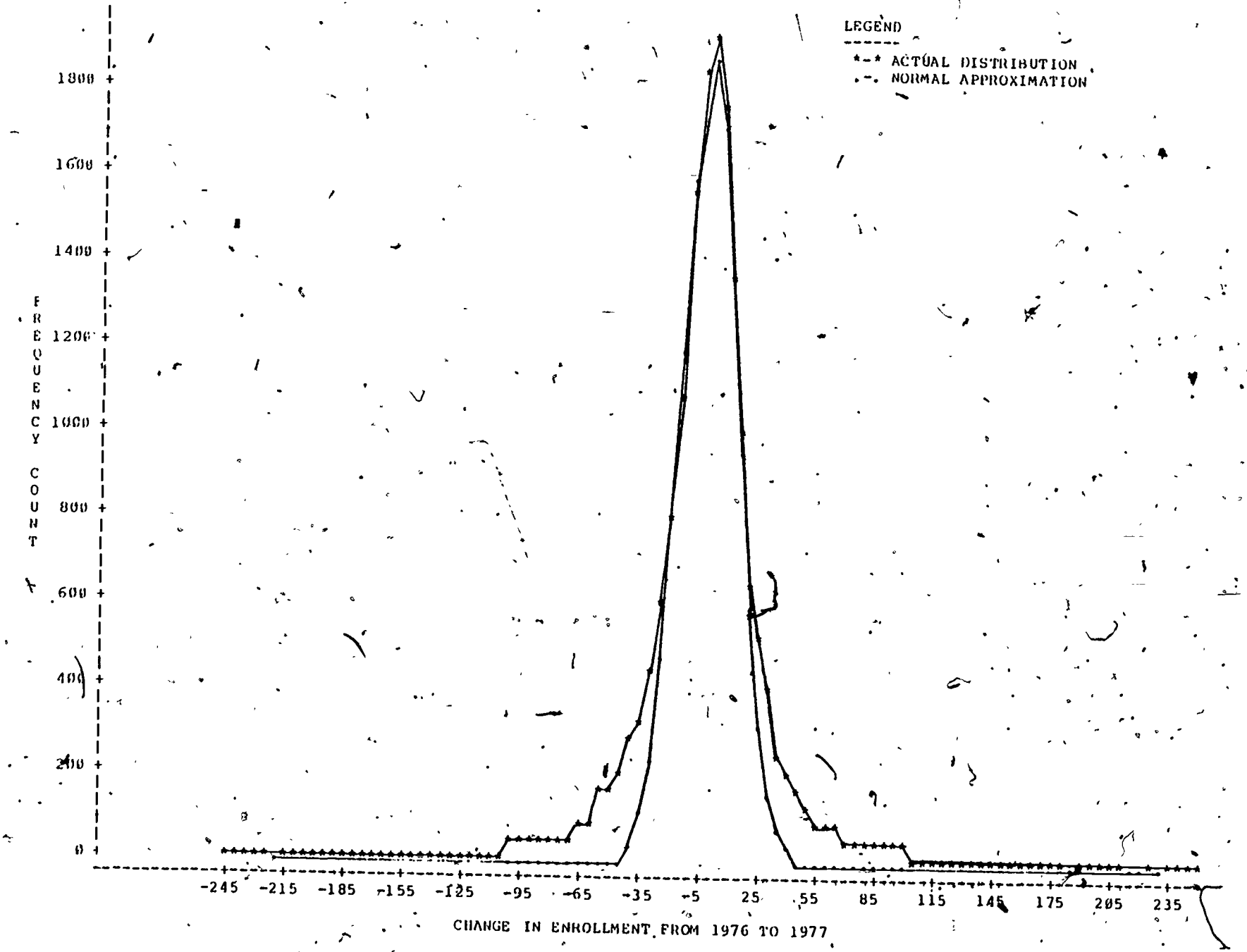

FIGURE 2. DISTRIBUTION OF 1977 ENROLLMENT CHANGES FOR NONPUBLIC SCHOOLS

thicker tails of the observed distribution could be due to a greater proportion of errors among the more extreme differences. This is, of course, very tentative. It is not necessary to estimate the error rate ahead of time unless it is desired to perform some form of cost-benefit analysis to determine an "optimal" cutoff point.

One problem in "fitting" a normal model to estimate the error rate is that the correct and error values are initially indistinguishable. If the overall standard deviation is used to estimate the standard deviation of the correct values, the resultant estimate will be too high by some unknown amount since error values have an additional variance component. As a result we will estimate that more of the extreme values are valid than is actually the case. In a recent study based on SAGE work, Fingerman (1981) showed that if the standard deviation of the correct values is estimated from the interquartile distance (actually as .74 times the distance from the first to the third quartile point), the resultant estimate is quite accurate, even where the proportion of errors is relatively large. The interquartile distance is influenced by the number of extreme cases but not by their degree of extremity while the usual variance estimator is strongly influenced by the degree of extremity of the most deviant cases. In a Monte Carlo simulation Fingerman found that when the variance of the distribution of "error" cases was nine times the variance of the distribution of valid responses and 10% of the cases were in error, the usual variance estimate based on all cases was 2.9 times too large, but the estimate based on the interquartile distance was only 1.1 times too large. Further, the estimate based on the interquartile distance was quite stable. The variance of the interquartile "variance" estimate was only 2% of the actual value compared to 30%, for the usual estimate based on all cases.

UNIVERSE COVERAGE

For much of the work that NCES does, estimates of totals, such as the total number of pupils, schools, teachers, and expenditures, are critical. For this reason, the issue of whether the universe has been fully covered is of particular concern. (If we were only estimating means, omitting some schools from the sampling or survey frame might not introduce serious bias, but if we want to know the total number of students such an omission will necessarily result in an undercount.)

One area of SAGE effort where the issue of coverage was of critical concern was in our work with the Nonpublic Elementary and Secondary School Surveys (McLaughlin & Wise, 1980). We began with a file of just over 18,000 schools from the 1977-1978 survey and merged these with a somewhat smaller number of schools from the 1976-77 survey. (The 76/77 files did not include nonrespondents.) The merging process was complicated by the fact that there was not a common identifier so that fallible name and address data had to be used to match schools. The process turned up the fact that both files contained some duplicate schools with small variation in the names and/or addresses. More importantly, each file contained a number of schools that were not on the other file. A sample of these schools were contacted and it was found that most of them were in fact operating both years. Other special cases were also identified, such as the fact that Mormon schools only reported aggregate data for the 1977-78 survey.

In the end, after the addition of the 1978-79 survey data and similar checking on unmatched schools, the total number of schools identified and considered open during the 1977-78 school year was estimated to be over 20,000 (20073) rather than the 18,103 initially identified. Needless to say, this reflects an increase of over 10% in the estimated number of nonpublic schools as well as in estimates of the number of students and teachers in these schools. (Later checks of state directories by SAGE indicated an additional undercoverage of approximately 10% in schools, or 1 or 2% in enrollment.) A current SAGE effort is designed to test alternative field strategies for assessing the adequacy of coverage in universe surveys such as this.

IMPUTATION OF MISSING DATA

The most ambitious SAGE effort in the area of survey data enhancement concerns imputation of missing data. This effort combined work on NCES's nonpublic school surveys with a general methodological development task to study procedures imputing missing values. Separate procedures were developed for imputing discrete (nominal) and continuous (interval) variables with or without prior year's data. Each of the final procedures was subjected to a special validation study where known values were masked and run through the imputation procedure. The real and imputed values were

compared to assess the extent of bias in estimates of means, variances, or relationships generated from the imputed values. The results of this validation were quite promising. The overall mean bias (due to missing data) in estimates generated from the final data was estimated to be less than one-half percent. Variances and relationships (correlations or conditional frequencies) were also reproduced reasonably well. The results were far and away superior to the two "easy" options for dealing with missing data--ignoring it or substituting mean values.

These results apply to the final procedures. During the course of this work, we learned the hard way about a number of pitfalls in the application of a regression approach to the imputation of missing values. These experiences were valuable for our subsequent work on general algorithm for the imputation of missing data. That work incorporated solutions to some of the sticky problems that we encountered, including the following. These problems illustrate the difficulty of avoiding serious bias in the values.

Variables with nonnormal distributions. Most of the continuous variables in this survey had strongly skewed distributions with no negative values and a small number of very large values. This was particularly true for the expenditure data. The regression approach occasionally gave predicted values that were negative. More frequently, when we went to add a random component reflecting the prediction error (to avoid shrinking the variance of the imputed values relative to the appropriate level), the random component caused the imputed value to become negative. In order to avoid having negative values (e.g., for enrollment) on the file, small positive values were substituted for the negative values. This, of course, led to a positive bias so that we had to introduce a corresponding truncation of relatively large values in order to compensate for the correction of negative values. This procedure is clearly unacceptable in general and is a strong rationale for use of some form of "hot deck" procedure that limits imputed values to the range of actually observed values instead of a formula procedure such as regression.

Problems with the use of derived variables. In predicting missing values from prior year's data, we were actually predicting the percent increase from other variables and then multiplying the prior value by the predicted rate of increase. Unfortunately, this led to another bias since

the expected value of the product of two random variables (the prior value times the rate of increase) is greater than the product of their expected values. Here too a correction was developed that proved satisfactory for each particular case. Initially, we had an even more severe problem in that we attempted to predict the log of the expenditure rate rather than the rate itself. This made sense because the expenditure data showed a somewhat lograthmic relationship to the potential predictors. It proved to be a disaster, however, since very small overestimates of the log led to rather large overestimates of the expenditure rate itself, so that when we converted back to real dollars, we had serious overestimates.

Preserving relationships among imputed values. A third sticky problem that surfaced was the difficulty of preserving true relationships among imputed values. For many nonresponding schools, very little was known, so that most of the values were imputed. If each missing value was imputed independently from the available values, relationships between the missing values would have been missed. For example, we imputed whether the school served boys or girls or both and whether the school included boarding students separately from the schools religous affiliation. Table 1 shows data from the validation study comparing the actual and imputed values. The actual values indicate that schools that served girls only were much less likely to include boarding students relative to other schools. This relationship was not found among the imputed values.

After having spent months developing tailor-made procedures for imputing missing values in the nonpublic school surveys, we sought to create an algorithm that would allow researchers to perform the equivalent work in an afternoon. The result of this effort was PROC IMPUTE (Wise & McLaughlin, 1980), a new procedure added to the Statistical Analysis System (SAS). By incorporating our algorithm into an existing statistical package, we eliminated the need for a researcher to duplicate efforts already spent defining variables, labels, missing data codes, etc. We also made the procedure more powerful in that it could be combined with the great flexibility already available in the SAS system for taking samples of cases, recoding variables, merging in additional data, and saving intermediate files.

The basic approach used in PROC IMPUTE is that a regression equation is developed for each variable with any missing values. For each equation, a two-way table giving the frequency of the actual values by the

Table 1

Actual and Imputed Relationship
between sex served and Boarding Facilities

Sex Served	Day Student Only		Some Boarding Students	
	Actual %	Imputed %	Actual %	Imputed %
Males Only	54.8	89.4	45.2	10.6
Females Only	83.1	89.2	16.9	10.8
Coed	94.1	91.2	5.9	8.8

predicted, regression function, values (divided into discrete categories) is developed. Figure 3 illustrates such a contingency table. For each missing value, a "predicted" value is generated using the regression function and then an "actual" value is selected randomly with probability proportional to the frequencies in the row of the two-way table corresponding to the predicted value. By using this procedure instead of just using the predicted values, we are certain that only values that actually occur are selected as imputed values, and we assure an appropriate variation for the imputed values.

One other feature of PROC IMPUTE is that the regression equations are developed in a "stepwise" manner. The first variable is imputed only from variables with no missing values. Each succeeding variable includes the variables already imputed as potential predictors so that imputed values are used in imputing other missing values. This is a significant difference from the BMDP procedure where only nonmissing values are used as predictors. After each missing value has been imputed, the procedure generates a second equation for "reimputing" each variable with missing values from all other variables. In practice, this second imputation is performed only if variables that were excluded in the initial imputation (because there came later in the initial list) had a significant correlation with the variable being imputed after partialling out the predictors that were used. In this way any significant relationships between variables with missing values are preserved, since each is used in the prediction of the other. A special procedure was developed to select an optimal ordering of the variables for the initial imputations. This procedure performs a "simultaneous" step-wise regression for all variables with missing values. At each step, a target variable and a new predictor variable are chosen that maximally reduce the uncertainty in the remaining missing values subject to the constraints imposed by the existing partial ordering of the variables. The pair selected then adds a new order constraint, that the predictor must precede the target variable in the imputation list. The process is continued until no more significant predictors are available.

Table 2 shows some results of a Monte Carlo study comparing the results of PROC IMPUTE to the results of the BMDP procedure. The results show that PROC IMPUTE was indeed successful at reproducing variances and

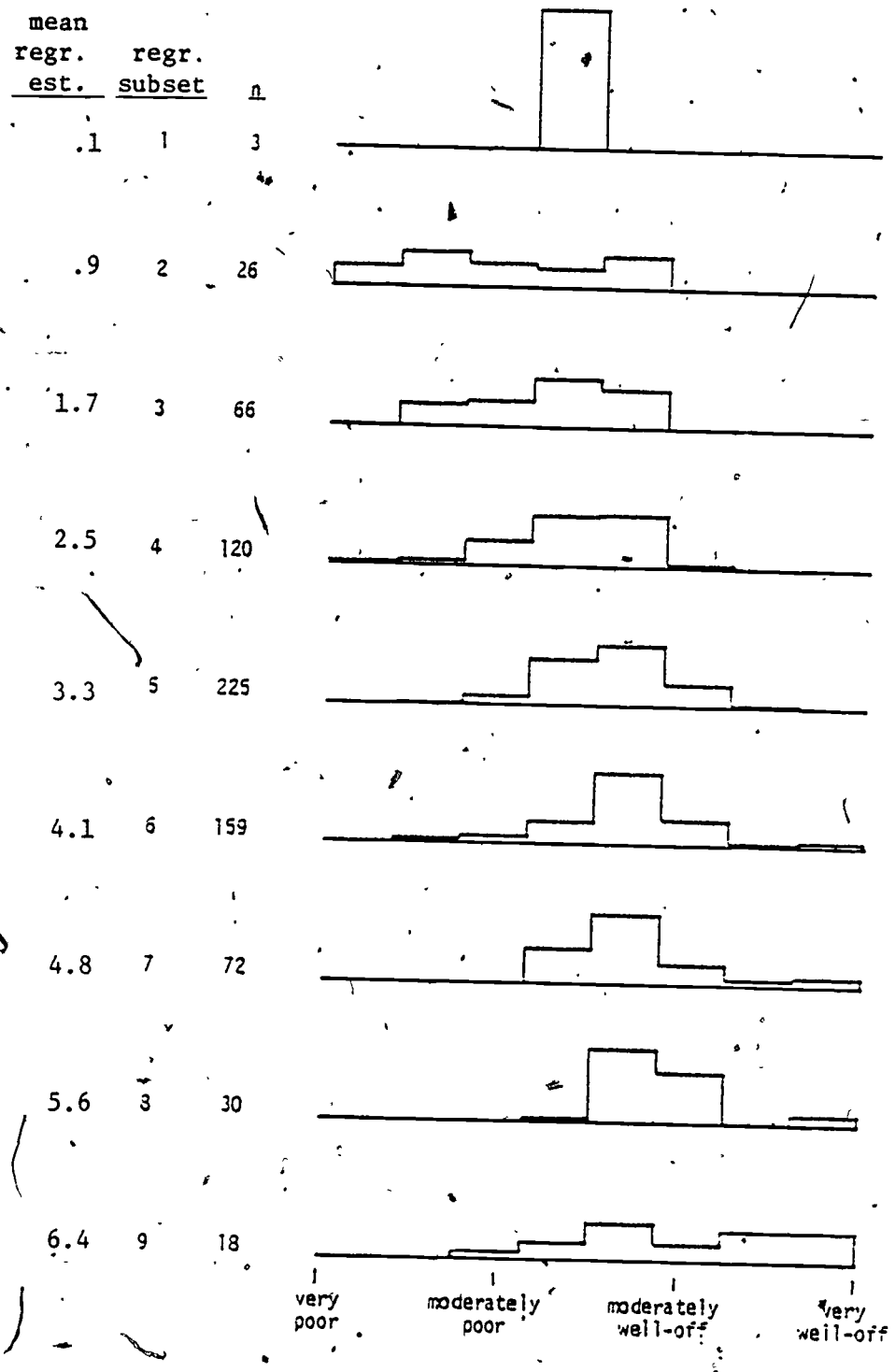

FIGURE 3. Distribution of target variable for each regression-function subset.

SCHOOL DISTRICT SES MEASURE
(School TV Utilization Study: NCES, 1979)

Note: The regression function was selected to account for maximum variance in the SES measure. Values were then partitioned into 9 discrete categories. The "n" refers to the number of cases in each regression-function category.

Table 2

Processing Time and Accuracy of Different BMDPAM Options and PROC IMPUTE

	Processing Time*	Error of Mean Estimate**	S.D. Estimate**	Error of Correlation Estimate***
BMDPAM Options:				
Mean Substitution	3.8	.549	-	.33
Singel Variable	5.6	.403	.558	-
Two Step	6.8	.400	.527	-
Total Regression	11.8	.392	.501	-
Stepwise Regression	25.6	.390	.508	.21
PROC IMPUTE	8.2	.383	.105	.15

* For a file with 20 variables and 1,000 observations. The processing time is in CPU seconds for an IBM 370/168 running under MVS.

** Average absolute error across 20 variables expressed in standard deviation units.

*** Root mean square errors averaged across all pairs of variables and all replications.

correlations while not sacrificing much in the accuracy of mean predictions. Copies of this procedure and instructions for setting up an appropriate SAS library can be obtained from AIR at cost.

SUMMARY

During the past two years SAGE has worked on the enhancement of survey data as one of its main themes. The work described here on the use of longitudinal merges for enhancing edits and for improving universe coverage and on the development of missing data imputation procedures that can be applied to a wide range of surveys. The current SAGE team is continuing work in the area of survey data enhancement including the development of survey error profiles and the study of appropriate analytic techniques.

REFERENCES

Fingerman, P. W. Robust measures of scale in outlier defection. Paper presented at the Sixth Annual SAS Users Group Conference, 1981.

McLaughlin, D. H., & Wise, L. L. Nonpublic Education of the Nation's Children (SAGE Technical Report No. 9). Palo Alto, CA: American Institutes for Research, 1980.

Wise, L. L., & McLaughlin, D. H. Guidebook for Imputation of Missing Data (SAGE Technical Report No. 17). Palo Alto, CA: American Institutes for Research, 1980.